Step size adaptation in first-order method for stochastic strongly convex programming

Peng Cheng
pc175@uow.edu.au

Abstract

We propose a first-order method for stochastic strongly convex optimization that attains $O(1/n)$ rate of convergence, analysis show that the proposed method is simple, easily to implement, and in worst case, asymptotically four times faster than its peers. We derive this method from several intuitive observations that are generalized from existing first order optimization methods.

I. Problem Setting

In this article we seek a numerical algorithm that iteratively approximates the solution w^* of the following strongly convex optimization problem:

$$w^* = \arg\min_{\Gamma_f} f(.) \tag{1}$$

where $f(.) : \Gamma_f \to \mathcal{R}$ is an unknown, not necessarily smooth, multivariate and λ-strongly convex function, with Γ_f its convex definition domain. The algorithm is not allowed to accurately sample $f(.)$ by any means since $f(.)$ itself is unknown. Instead the algorithm can call stochastic oracles $\tilde{\omega}(.)$ at chosen points $\tilde{x}_1, \dots, \tilde{x}_n$, which are unbiased and independent probabilistic estimators of the first-order local information of $f(.)$ in the vicinity of each x_i:

$$\tilde{\omega}(x_i) = \{\tilde{f}_i(x_i), \bigtriangledown \tilde{f}_i(x_i)\} \tag{2}$$

where $\bigtriangledown$ denotes random subgradient operator, $\tilde{f}_i(.) : \Gamma_f \to \mathcal{R}$ are independent and identically distributed (i.i.d.) functions that satisfy:

$$\text{(unbiased) } \mathbb{E}[\tilde{f}_i(.)] = f(.) \quad \forall i \tag{3a}$$

$$\text{(i.i.d) } \mathrm{Cov}\left(\tilde{f}_i(.), \tilde{f}_j(.)\right) = 0 \quad \forall i \neq j \tag{3b}$$

Solvers to this kind of problem are highly demanded by scientists in large scale computational learning, in which the first-order stochastic oracle is the only measurable information of $f(.)$ that scale well with both dimensionality and scale of the learning problem. For example, a stochastic first-order oracle in structural risk minimization (a.k.a. training a support vector machine) can be readily obtained in $O(1)$ time [1].

II. Algorithm

The proposed algorithm itself is quite simple but with a deep proof of convergence. The only improvement comparing to SGD is the selection of step size in each iteration, which however, results in substantial boost of performance, as will be shown in the next section.

III. Analysis

The proposed algorithm is designed to generate an output y that reduces the suboptimality:

$$S(y) = f(y) - \min f(.) \tag{4}$$

as fast as possible after a number of operations. We derive the algorithm by several intuitive observations that are generalized from existing first order methods. First, we start from worst-case upper-bounds of $S(y)$ in deterministic programming:

Lemma 1. *(Cutting-plane bound [4]): Given n deterministic oracles $\Omega_n = \{\omega(x_1), \dots, \omega(x_n)\}$ defined by:*

$$\omega(x_i) = \{f(x_i), \bigtriangledown f(x_i)\} \tag{5}$$

If $f(.)$ is a λ-strongly convex function, then $\min f(.)$ is unimprovably lower bounded by:

$$\min f(.) \geq \max_{i=1\dots n} p_i(w^*) \geq \min \max_{i=1\dots n} p_i(.) \tag{6}$$

Algorithm 1

```
Receive $x_1, \Gamma_f, \lambda$
$u_1 \leftarrow 1$, $y_1 \leftarrow x_1$
Receive $\tilde{f}_1(x_1), \bigtriangledown \tilde{f}_1(x_1)$
$\tilde{P}_1(.) \leftarrow \Gamma_f \left\{ \tilde{f}_1(x_1) + \langle \bigtriangledown \tilde{f}_1(x_1), . - x_i \rangle + \frac{\lambda}{2}||. - x_1||^2 \right\}$
for $i = 2, \ldots, n$ do
    $x_i \leftarrow \arg\min \tilde{P}_{i-1}(.)$
    Receive $\tilde{f}_i(x_i), \bigtriangledown \tilde{f}_i(x_i)$
    $\tilde{p}_i(.) \leftarrow \Gamma_f \left\{ \tilde{f}_i(x_i) + \langle \bigtriangledown \tilde{f}_i(x_i), . - x_i \rangle + \frac{\lambda}{2}||. - x_i||^2 \right\}$
    $\tilde{P}_i(.) \leftarrow (1 - \frac{u_{i-1}}{2})\tilde{P}_{i-1}(.) + \frac{u_{i-1}}{2}\tilde{p}_i(.)$
    $y_i \leftarrow (1 - \frac{u_{i-1}}{2})y_{i-1} + \frac{u_{i-1}}{2}x_i$
    $u_i \leftarrow u_{i-1} - \frac{u_{i-1}^2}{4}$
end for
Output $y_n$
```

where w^* is the unknown minimizer defined by (1)*, and $p_i(.) : \Gamma_f \rightarrow \mathcal{R}$ are proximity control functions (or simply prox-functions) defined by $p_i(.) = f(x_i) + \langle \bigtriangledown f(x_i), . - x_i \rangle + \frac{\lambda}{2}||. - x_i||^2$.*

Proof:

By strong convexity of $f(.)$ we have:

$$\begin{aligned}
& & \mathbb{B}_f(.||x_i) \geq \frac{\lambda}{2}||. - x_i||^2 & \\
\Longrightarrow & & f(.) \geq p_i(.) & \\
\Longrightarrow & & f(.) \geq \max_{i=1,\ldots,n} p_i(.) & \qquad (7) \\
\Longrightarrow & & \min f(.) = f(w^*) \geq \max_{i=1,\ldots,n} p_i(w^*) \geq \min \max_{i=1,\ldots,n} p_i(.) & \qquad (8)
\end{aligned}$$

where $\mathbb{B}_f(x_1||x_2) = f(x_1) - f(x_2) - \langle \bigtriangledown f(x_2), x_1 - x_2 \rangle$ denotes the Bregman divergence between two points $x_1, x_2 \in \Gamma_f$. Both sides of (7) and (8) become equal if $f(.) = \max_{i=1,\ldots,n} p_i(.)$, so this bound cannot be improved without any extra condition.

■

Lemma 2. *(Jensen's inequality for strongly convex function) Given n deterministic oracles $\Omega_n = \{\omega(x_1), \ldots, \omega(x_n)\}$ defined by* (5)*. If $f(.)$ is a λ-strongly convex function, then for all $\alpha_1, \ldots, \alpha_n$ that satisfy $\sum_{i=1}^n \alpha_i = 1, \alpha_i \geq 0 \quad \forall i$, $f(y)$ is unimprovably upper bounded by:*

$$f(y) \leq \sum_{i=1}^n \alpha_i f(x_i) - \frac{\lambda}{2}\sum_{i=1}^n \alpha_i ||x_i - y||^2 \qquad (9)$$

where $y = \sum_{i=1}^n \alpha_i x_i$.

Proof:

By strong convexity of $f(.)$ we have:

$$\begin{aligned}
& & \mathbb{B}_f(x_i||y) &\geq \frac{\lambda}{2}||x_i - y||^2 \\
\Longrightarrow & & f(y) &\leq f(x_i) - \langle \bigtriangledown f(y), x_i - y \rangle - \frac{\lambda}{2}||x_i - y||^2 \\
\Longrightarrow & & f(y) &\leq \sum_{i=1}^n \alpha_i f(x_i) - \langle \bigtriangledown f(y), \sum_{i=1}^n \alpha_i x_i - y \rangle - \frac{\lambda}{2}\sum_{i=1}^n \alpha_i ||x_i - y||^2 \\
& & &\leq \sum_{i=1}^n \alpha_i f(x_i) - \frac{\lambda}{2}\sum_{i=1}^n \alpha_i ||x_i - y||^2
\end{aligned}$$

Both sides of all above inequalities become equal if $f(.) = \frac{\lambda}{2}||. - y||^2 + \langle c_1, . \rangle + c_2$, where c_1 and c_2 are constants, so this bound cannot be improved without any extra condition.

■

Immediately, the optimal A that yields the lowest upper bound of $f(y)$ can be given by:

$$A = \arg \min_{\substack{\sum_{i=1}^n \alpha_i = 1 \\ \alpha_i \geq 0 \forall i}} \left\{ \sum_{i=1}^n \alpha_i f(x_i) - \frac{\lambda}{2} \sum_{i=1}^n \alpha_i ||x_i - \sum_{j=1}^n \alpha_j x_j||^2 \right\} \tag{10}$$

Combining with (4), (6), we have an deterministic upper bound of $S(y)$:

$$S(y) \leq \min_{\substack{\sum_{i=1}^n \alpha_i = 1 \\ \alpha_i \geq 0 \forall i}} \left\{ \sum_{i=1}^n \alpha_i f(x_i) - \frac{\lambda}{2} \sum_{i=1}^n \alpha_i ||x_i - \sum_{j=1}^n \alpha_j x_j||^2 \right\} - \max_{i=1 \dots n} p_i(w^*) \tag{11}$$

This bound is quite useless at the moment as we are only interested in bounds in stochastic programming. The next lemma will show how it can be generalized in later case.

Lemma 3. *Given n stochastic oracles $\tilde{\Omega}_n = \{\tilde{\omega}(x_1), \dots, \tilde{\omega}(x_n)\}$ defined by* (2)*, if $y(., \dots, .) : \mathcal{H}^n \times \Gamma_f^n \to \Gamma_f$ and $U(., \dots, .) : \mathcal{H}^n \times \Gamma_f^n \to \mathcal{R}$ are functionals of $\tilde{f}_i(.)$ and x_i that satisfy:*

$$U(f, \dots, f, x_1, \dots, x_n) \geq S(y(f, \dots, f, x_1, \dots, x_n)) \tag{12a}$$

$$U(\tilde{f}_1, \dots, \tilde{f}_n, x_1, \dots, x_n) \text{ is convex w.r.t. } \tilde{f}_1, \dots, \tilde{f}_n \tag{12b}$$

$$\mathbb{E}[\langle \bigtriangledown_{f,\dots,f} U(f, \dots, f, x_1, \dots, x_n), [\tilde{f}_1 - f, \dots, \tilde{f}_n - f]^T \rangle] \leq 0 \tag{12c}$$

then $\mathbb{E}[S(y(f, \dots, f, x_1, \dots, x_n))]$ is upper bounded by $U(\tilde{f}_1, \dots, \tilde{f}_n, x_1, \dots, x_n)$.

Proof:

Assuming that $\delta_i(.) : \Gamma_f \to \mathcal{R}$ are perturbation functions defined by

$$\delta_i(.) = \tilde{f}_i(.) - f(.) \tag{13}$$

we have:

$$\begin{aligned} U(\tilde{f}_{1,\dots,n}, x_{1,\dots,n}) &\geq U(f + \delta_1, \dots, f + \delta_n, x_{1,\dots,n}) \\ \text{(by (12b))} &= U(f, \dots, f, x_{1,\dots,n}) + \langle \bigtriangledown_{f,\dots,f} U(f, \dots, f, x_{1,\dots,n}), [\delta_{1,\dots,n}]^T \rangle \\ \text{(by (12a))} &\geq S(y(f, \dots, f, x_{1,\dots,n})) + \langle \bigtriangledown_{f,\dots,f} U(f, \dots, f, x_{1,\dots,n}), [\delta_{1,\dots,n}]^T \rangle \end{aligned} \tag{14}$$

Moving δ_i to the left side:

$$\begin{aligned} \mathbb{E}[S(y(f, \dots, f, x_{1,\dots,n}))] &\leq U(\tilde{f}_{1,\dots,n}, x_{1,\dots,n}) + \mathbb{E}[\langle \bigtriangledown_{f,\dots,f} U(f, \dots, f, x_{1,\dots,n}), [\delta_{1,\dots,n}]^T \rangle] \\ \text{(by (12c))} &\leq U(\tilde{f}_{1,\dots,n}, x_{1,\dots,n}) \end{aligned}$$

■

Clearly, according to (12b), setting:

$$U(\tilde{f}_{1,\dots,n}, x_{1,\dots,n}) = \min_{\substack{\sum_{i=1}^n \alpha_i = 1 \\ \alpha_i \geq 0 \forall i}} \left\{ \sum_{i=1}^n \alpha_i \tilde{f}_i(x_i) - \frac{\lambda}{2} \sum_{i=1}^n \alpha_i ||x_i - \sum_{j=1}^n \alpha_j x_j||^2 \right\} - \max_{i=1 \dots n} \tilde{p}_i(w^*) \tag{15}$$

by substituting $f(.)$ and $p_i(.)$ in (11) respectively with $\tilde{f}_i(.)$ defined by (3) and $\tilde{p}_i(.) : \Gamma_f \to \mathcal{R}$ defined by:

$$\tilde{p}_i(.) = \tilde{f}_i(x_i) + \langle \bigtriangledown \tilde{f}_i(x_i), . - x_i \rangle + \frac{\lambda}{2} ||. - x_i||^2 \tag{16}$$

is not an option, because $\min_{\substack{\sum_{i=1}^n \alpha_i = 1 \\ \alpha_i \geq 0 \forall i}} \{.\}$ and $- \max_{i=1 \dots n} \{.\}$ are both concave, $\sum_{i=1}^n \alpha_i \tilde{f}_i(x_i)$ and $\tilde{p}_i(w^*)$ are both linear to $\tilde{f}_i(.)$, and $\frac{\lambda}{2} \sum_{i=1}^n \alpha_i ||x_i - \sum_{j=1}^n \alpha_j x_j||^2$ is irrelevant to $\tilde{f}_i(.)$. This prevents asymptotically fast cutting-plane/bundle methods [4], [8], [2] from being directly applied on stochastic oracles without any loss of performance. As a result, to decrease (4) and satisfy (12b) our options boil down to replacing $\min_{\substack{\sum_{i=1}^n \alpha_i=1 \\ \alpha_i \geq 0 \forall i}}\{.\}$ and $-\max_{i=1\dots n}\{.\}$ in (15) with their respective lowest convex upper bound:

$$\begin{aligned} U_{(A,B)}(\tilde{f}_{1,\dots,n}, x_{1,\dots,n}) &= \sum_{i=1}^n \alpha_i \tilde{f}_i(x_i) - \frac{\lambda}{2}\sum_{i=1}^n \alpha_i ||x_i - \sum_{j=1}^n \alpha_j x_j||^2 - \sum_{i=1}^n \beta_i \tilde{p}_i(w^*) \\ &= \sum_{i=1}^n \alpha_i \tilde{f}_i(x_i) - \frac{\lambda}{2}\sum_{i=1}^n \alpha_i ||x_i - \sum_{j=1}^n \alpha_j x_j||^2 - \tilde{P}_n(w^*) \end{aligned} \tag{17}$$

where $\tilde{P}_n(.) : \Gamma_f \rightarrow \mathcal{R}$ is defined by:

$$\tilde{P}_n(.) = \sum_{i=1}^n \beta_i \tilde{p}_i(.) \tag{18}$$

and $A = [\alpha_1, \dots, \alpha_n]^T$, $B = [\beta_1, \dots, \beta_n]^T$ are constant n-dimensional vectors, with each α_i, β_i satisfying:

$$\sum_{i=1}^n \alpha_i = 1 \qquad\qquad \alpha_i \geq 0 \quad \forall i \tag{19a}$$

$$\sum_{i=1}^n \beta_i = 1 \qquad\qquad \beta_i \geq 0 \quad \forall i \tag{19b}$$

accordingly $y((\tilde{f}_1, \dots, \tilde{f}_n, x_1, \dots, x_n)$ can be set to:

$$y_{(A,B)}(x_1, \dots, x_n) = \sum_{i=1}^n \alpha_i x_i \tag{20}$$

such that (12a) is guaranteed by lemma 2. It should be noted that A and B must both be constant vectors that are independent from all stochastic variables, otherwise the convexity condition (12b) may be lost. For example, if we always set β_i as the solution of the following problem:

$$\beta_i = \arg \max_{\substack{\sum_{i=1}^n \beta_i=1 \\ \beta_i \geq 0 \forall i}} \left\{ \sum_{i=1}^n \beta_i \tilde{p}_i(w^*) \right\}$$

then $\tilde{P}_n(w^*)$ will be no different from the cutting-plane bound (6). Finally, (12c) can be validated directly by substituting (17) back into (12c):

$$\langle \bigtriangledown_{f,\dots,f} U_{(A,B)}(f, \dots, f, x_{1,\dots,n}), [\delta_{1,\dots,n}]^T \rangle = \sum_{i=1}^n [(\alpha_i - \beta_i)\delta_i(x_i) - \langle \bigtriangledown \delta_i(x_i), \beta_i(w^* - x_i) \rangle] \tag{21}$$

Clearly $\mathbb{E}[(\alpha_i - \beta_i)\delta_i(x_i)] = 0$ and $\mathbb{E}[\langle \bigtriangledown \delta_i(x_i), w^* \rangle] = 0$ can be easily satisfied because α_i and β_i are already set to constants to enforce (12b), and by definition $w^* = \arg\min f(.)$ is a deterministic (yet unknown) variable in our problem setting, while both $\delta_i(x_i)$ and $\bigtriangledown \delta_i(x_i)$ are unbiased according to (3a). Bounding $\mathbb{E}[\langle \bigtriangledown \delta_i(x_i), x_i \rangle]$ is a bit harder but still possible: In all optimization algorithms, each x_i can either be a constant, or chosen from Γ_f based on previous $\tilde{f}_1(.), \dots, \tilde{f}_{i-1}(.)$ (x_i cannot be based on $\tilde{f}_i(.)$ that is still unknown by the time x_i is chosen). By the i.i.d. condition (3b), they are all independent from $\tilde{f}_i(.)$, which implies that x_i is also independent from $\tilde{f}_i(x_i)$:

$$\mathbb{E}[\langle \bigtriangledown \delta_i(x_i), x_i \rangle] = 0 \tag{22}$$

As a result, we conclude that (21) satisfies $\mathbb{E}[\langle \bigtriangledown_{f,\dots,f} U_{(A,B)}(f, \dots, f, x_{1,\dots,n}), [\delta_{1,\dots,n}]^T \rangle] = 0$, and subsequently $U_{(A,B)}$ defined by (17) satisfies all three conditions of Lemma 3. At this point we may construct an algorithm that uniformly reduces $\max_{w^*} U_{(A,B)}$ by iteratively calling new stochastic oracles and updating A and B. Our main result is summarized in the following theorem:

Theorem 1. *For all λ-strongly convex function $F(.)$, assuming that at some stage of an algorithm, n stochastic oracles $\tilde{\omega}(x_1), \dots, \tilde{\omega}(x_n)$ have been called to yield a point $y_{(A^n,B^n)}$ defined by* (20) *and an upper bound $\hat{U}_{(A^n,B^n)}$ defined by:*

$$\hat{U}_{(A^n,B^n)}(\tilde{f}_{1,\dots,n}, x_{1,\dots,n}) = U_{(A^n,B^n)}(\tilde{f}_{1,\dots,n}, x_{1,\dots,n}) + \frac{\lambda}{2}\sum_{i=1}^{n} \alpha_i ||x_i - \sum_{j=1}^{n} \alpha_j x_j||^2 + \left(\tilde{P}_n(w^*) - \min \tilde{P}_n(.)\right)$$
$$= \sum_{i=1}^{n} \alpha_i \tilde{f}_i(x_i) - \min \tilde{P}_n(.) \tag{23}$$

where A^n and B^n are constant vectors satisfying (19) *(here ^{n} in A^n and B^n denote superscripts, should not be confused with exponential index), if the algorithm calls another stochastic oracle $\tilde{\omega}(x_{n+1})$ at a new point x_{n+1} given by:*

$$x_{n+1} = \arg\min \tilde{P}_n(.)$$

and update A and B by:

$$A^{n+1} = \left[\left(1 - \frac{\lambda}{G^2}\hat{U}_{(A^n,B^n)}\right)(A^n)^T, \frac{\lambda}{G^2}\hat{U}_{(A^n,B^n)}\right]^T \tag{24a}$$
$$B^{n+1} = \left[\left(1 - \frac{\lambda}{G^2}\hat{U}_{(A^n,B^n)}\right)(B^n)^T, \frac{\lambda}{G^2}\hat{U}_{(A^n,B^n)}\right]^T \tag{24b}$$

where $G = \max || \bigtriangledown \tilde{f}_i(.)||$, then $\hat{U}_{A^{n+1},B^{n+1}}$ is bounded by:

$$\hat{U}_{(A^{n+1},B^{n+1})} \leq \hat{U}_{(A^n,B^n)} - \frac{\lambda}{2G^2}\hat{U}^2_{(A^n,B^n)} \tag{25}$$

Proof:

First, optimizing and caching all elements of A^n or B^n takes at least $O(n)$ time and space, which is not possible in large scale problems. So we confine our options of A^{n+1} and B^{n+1} to setting:

$$[\alpha_1^{n+1}, \dots, \alpha_n^{n+1}]^T = (1 - \alpha_{n+1}^{n+1})[\alpha_1^n, \dots, \alpha_n^n]^T \tag{26a}$$
$$[\beta_1^{n+1}, \dots, \beta_n^{n+1}]^T = (1 - \beta_{n+1}^{n+1})[\beta_1^n, \dots, \beta_n^n]^T \tag{26b}$$

such that previous $\sum_{i=1}^{n} \alpha_i \tilde{F}(x_i)$ and $\sum_{i=1}^{n} \beta_i \tilde{p}_i(.)$ can be summed up in previous iterations in order to produce a 1-memory algorithm instead of an ∞-memory one, without violating (19). Consequently $\hat{U}_{(A^{n+1},B^{n+1})}$ can be decomposed into:

$$\begin{aligned}
\hat{U}_{(A^{n+1},B^{n+1})} &= \sum_{i=1}^{n+1} \alpha_i^{n+1} \tilde{f}(x_i) - \min \tilde{P}_{n+1}(.) \\
\text{(by (16), (18), (19))} &\leq \sum_{i=1}^{n+1} \alpha_i^{n+1} \tilde{f}(x_i) - \left[\sum_{i=1}^{n+1} \beta_i^{n+1} \tilde{p}_i(\tilde{x}_n^*) - \frac{1}{2\lambda} || \bigtriangledown \sum_{i=1}^{n+1} \beta_i^{n+1} \tilde{p}_i(\tilde{x}_n^*)||^2\right] \\
\text{(by (26))} &= \left[(1 - \alpha_{n+1}^{n+1}) \sum_{i=1}^{n} \alpha_i^n \tilde{f}(x_i) - (1 - \beta_{n+1}^{n+1}) \sum_{i=1}^{n} \beta_i^n \tilde{p}_i(\tilde{x}_n^*)\right] \\
&\quad + \left[\alpha_{n+1}^{n+1} \tilde{f}(x_{n+1}) - \beta_{n+1}^{n+1} \tilde{p}_{n+1}(\tilde{x}_n^*)\right] + \frac{(\beta_{n+1}^{n+1})^2}{2\lambda} || \bigtriangledown \tilde{p}_{n+1}(\tilde{x}_n^*)||^2
\end{aligned}$$

where $\tilde{x}_n^* = \arg\min \tilde{P}_n(.)$, setting $a_{n+1}^{n+1} = b_{n+1}^{n+1}$ and $x_{n+1} = \tilde{x}_n^*$ eliminates the second term:

$$\begin{aligned}
\hat{U}_{(A^{n+1},B^{n+1})} &= (1 - \alpha_{n+1}^{n+1}) \left(\sum_{i=1}^{n} \alpha_i^n \tilde{f}(x_i) - \tilde{P}_n(\tilde{x}_n^*)\right) + \frac{(\alpha_{n+1}^{n+1})^2}{2\lambda} || \bigtriangledown \tilde{p}_{n+1}(x_{n+1})||^2 \\
\text{(by (16))} &= (1 - \alpha_{n+1}^{n+1}) \hat{U}_{(A^n,B^n)} + \frac{(\alpha_{n+1}^{n+1})^2}{2\lambda} || \bigtriangledown \tilde{f}_{n+1}(x_{n+1})||^2 \\
(G \geq || \bigtriangledown \tilde{f}_i(.)||) &\leq (1 - \alpha_{n+1}^{n+1}) \hat{U}_{(A^n,B^n)} + \frac{(\alpha_{n+1}^{n+1})^2 G^2}{2\lambda}
\end{aligned} \tag{27}$$

Let $u_i = \frac{2\lambda}{G^2}\hat{U}_{(A^i,B^i)}$, minimizing the right side of (27) over α_{n+1}^{n+1} yields:

$$\alpha_{n+1}^{n+1} = \arg\min_{\alpha}\{(1 - \alpha)u_n + \alpha^2\} = \frac{u_n}{2} = \frac{\lambda}{G^2}\hat{U}_{(A^n,B^n)} \tag{28}$$

In this case:

$$\begin{aligned} & u_{n+1} \leq u_n - \frac{u_n^2}{4} \qquad (29) \\ \Longrightarrow \quad & \hat{U}_{(A^{n+1},B^{n+1})} \leq \hat{U}_{(A^n,B^n)} - \frac{2\lambda}{G^2}\hat{U}^2_{(A^n,B^n)} \end{aligned}$$

■

Given an arbitrary initial oracle $\tilde{\omega}(x_1)$ and apply the updating rule in theorem 1 recursively results in algorithm 1, accordingly we can prove its asymptotic behavior by induction:

Corollary 1. *The final point y_n obtained by applying algorithm 1 on arbitrary λ-strongly convex function $f(.)$ has the following worst-case rate of convergence:*

$$\mathbb{E}[f(y_n)] - \min f(.) \leq \frac{2G^2}{\lambda(n+3)}$$

Proof:

First, by (29) we have:

$$\frac{1}{u_{n+1}} \geq \frac{1}{u_n\left(1 - \frac{u_n}{4}\right)} = \frac{1}{u_n} + \frac{1}{4 - u_n} \geq \frac{1}{u_n} + \frac{1}{4} \qquad (30)$$

On the other hand, by strong convexity, for all $x_1 \in \Gamma_f$ we have:

$$f(x_1) - \min f(.) \leq \frac{1}{2\lambda}|| \bigtriangledown f(x_1)||^2 \leq \frac{G^2}{2\lambda} \qquad (31)$$

Setting $\hat{U}_{(1,1)} = \frac{G^2}{2\lambda}$ as intial condition and apply (30) recursively induces the following generative function:

$$\begin{aligned} & \frac{1}{u_n} \geq 1 + \frac{n-1}{4} = \frac{n+3}{4} \\ \Longrightarrow \quad & u_n \leq \frac{4}{n+3} \\ \Longrightarrow \quad & \hat{U}_{(A^n,B^n)} \leq \frac{2G^2}{\lambda(n+3)} \\ \Longrightarrow \quad & \mathbb{E}[f(y_n)] - \min f(.) \leq \frac{2G^2}{\lambda(n+3)} - \frac{\lambda}{2}\sum_{i=1}^{n} \alpha_i^n ||x_i - y_n||^2 - \left(\tilde{P}_n(w^*) - \min \tilde{P}_n(.)\right) \end{aligned}$$

■

This worst-case rate of convergence is four times faster than Epoch-GD ($\frac{8G^2}{\lambda n}$) [3], [5] or Cutting-plane/Bundle Method ($\frac{8G^2}{\lambda\left[n+2-\log_2\left(\frac{\lambda f(x_1)}{4G^2}\right)\right]}$) [4], [8], [2], and is indefinitely faster than SGD ($\frac{ln(n)G^2}{2\lambda n}$) [1], [6].

IV. High Probability Bound

An immediate result of Corollary 1 is the following high probability bound yielded by Markov inequality:

$$\Pr\left(S(y_n) \geq \frac{2G^2}{\lambda(n+3)\eta}\right) \leq \eta \qquad (32)$$

where $1 - \eta \in [0, 1]$ denotes the confidence of the result y_n to reach the desired suboptimality. In most cases (particularly when $\eta \approx 0$, as demanded by most applications) this bound is very loose and cannot demonstrate the true performance of the proposed algorithm. In this section we derive several high probability bounds that are much less sensitive to small η comparing to (32).

Corollary 2. *The final point y_n obtained by applying algorithm 1 on arbitrary λ-strongly convex function $F(.)$ has the following high probability bounds:*

$$\Pr\left(S(y_n) \geq t + \frac{2G^2}{\lambda(n+3)}\right) \leq \exp\left\{-\frac{t^2(n+2)}{16D^2\sigma^2}\right\} \tag{33a}$$

$$\Pr\left(S(y_n) \geq t + \frac{2G^2}{\lambda(n+3)}\right) \leq \frac{1}{2}\exp\left\{-\frac{t^2(n+2)}{8\tilde{G}^2D^2}\right\} \tag{33b}$$

$$\Pr\left(S(y_n) \geq t + \frac{2G^2}{\lambda(n+3)}\right) \leq \exp\left\{-\frac{t(n+2)}{4\tilde{G}D}\ln\left(1+\frac{t\tilde{G}}{2D\sigma^2}\right)\right\} \tag{33c}$$

where constants $\tilde{G} = \max ||\bigtriangledown\delta_i(.)||$, $\sigma^2 = \max \mathrm{Var}(\bigtriangledown\tilde{f}_i(.))$ are maximal range and variance of each stochastic subgradient respectively, and $D = \max_{x_1,x_2\in\Gamma_f} ||x_1 - x_2||$ is the largest distance between two points in Γ_f.

Proof:

We start by expanding the right side of (14), setting $A^n = B^n$ and substituting (21) back into (14) yields:

$$\begin{aligned} S(y_n) &\leq U_{(A^n,A^n)}(\tilde{f}_{1,\dots,n}, x_{1,\dots,n}) - \sum_{i=1}^{n}\alpha_i^n\langle\bigtriangledown\delta_i(x_i), x_i - w^*\rangle \\ \text{(by Corollary 1)} &\leq \frac{2G^2}{\lambda(n+3)} + \sum_{i=1}^{n}\alpha_i^n r_i \end{aligned} \tag{34}$$

with each $r_i = -\langle\bigtriangledown\delta_i(x_i), x_i - w^*\rangle$ satisfying:

$$\begin{aligned} \text{(Cauchy's inequality)} \; -||\bigtriangledown\delta_i(x_i)||||x_i - w^*|| &\leq r_i \leq ||\bigtriangledown\delta_i(x_i)||||x_i - w^*|| \\ -\tilde{G}D &\leq r_i \leq \tilde{G}D \end{aligned} \tag{35}$$

$$\begin{aligned} \mathrm{Var}(r_i) &= \mathbb{E}[(\langle\bigtriangledown\delta_i(x_i), x_i - w^*\rangle - \mathbb{E}[\langle\bigtriangledown\delta_i(x_i), x_i - w^*\rangle])^2] \\ \text{(by (22))} &= \mathbb{E}[(\langle\bigtriangledown\delta_i(x_i), x_i - w^*\rangle)^2] \\ \text{(Cauchy's inequality)} &\leq \mathbb{E}[||\bigtriangledown\delta_i(x_i)||^2||x_i - w^*||^2] \\ &\leq D^2\mathbb{E}[||\bigtriangledown\delta_i(x_i)||^2] = D^2\mathrm{Var}\left(\bigtriangledown\tilde{f}_i(x_i)\right) \leq D^2\sigma^2 \end{aligned} \tag{36}$$

This immediately expose $S_n(y_{(A^n,A^n)}(x_{1,\dots,n}))$ to several inequalities in non-parametric statistics that bound the probability of sum of independent random variables:

$$\begin{aligned} \text{(generalized Chernoff bound)} \Pr\left(\sum_{i=1}^{n}\alpha_i^n r_i \geq t\right) &\leq \exp\left\{-\frac{t^2}{4\mathrm{Var}\left(\sum_{i=1}^{n}\alpha_i^n r_i\right)}\right\} \\ \text{(by (36))} &\leq \exp\left\{-\frac{t^2}{4D^2\sigma^2\sum_{i=1}^{n}(\alpha_i^n)^2}\right\} \end{aligned} \tag{37a}$$

$$\begin{aligned} \text{(Azuma-Hoeffding inequality)} \Pr\left(\sum_{i=1}^{n}\alpha_i^n r_i \geq t\right) &\leq \frac{1}{2}\exp\left\{-\frac{2t^2}{\sum_{i=1}^{n}(\alpha_i^n)^2(\max r_i - \min r_i)^2}\right\} \\ \text{(by (35))} &\leq \frac{1}{2}\exp\left\{-\frac{2t^2}{4\tilde{G}^2D^2\sum_{i=1}^{n}(\alpha_i^n)^2}\right\} \end{aligned} \tag{37b}$$

$$\begin{aligned} \text{(Bennett inequality)} \Pr\left(\sum_{i=1}^{n}\alpha_i^n r_i \geq t\right) &\leq \exp\left\{-\frac{t}{2\max||\alpha_i^n\epsilon_i||}\ln\left(1+\frac{t\max||\alpha_i\epsilon_i||}{\mathrm{Var}\left(\sum_{i=1}^{n}\alpha_i^n r_i\right)}\right)\right\} \\ \text{(by (35), (36))} &\leq \exp\left\{-\frac{t}{2\tilde{G}D\max\alpha_i}\ln\left(1+\frac{t\tilde{G}\max\alpha_i^n}{D\sigma^2\sum_{i=1}^{n}(\alpha_i^n)^2}\right)\right\} \end{aligned} \tag{37c}$$

In case of algorithm 1, if A^n is recursively updated by (24), then each two consecutive α_i^n has the following property:

$$\begin{aligned}
&\text{(by (24))}\ \frac{\alpha_{i+1}^n}{\alpha_i^n} = \frac{\alpha_{i+1}^{i+1}}{\alpha_i^{i+1}} = \frac{\alpha_{i+1}^{i+1}}{\alpha_i^i(1-\alpha_{i+1}^{i+1})} \\
&\text{(by (28), (29))} = \frac{u_{i-1} - \frac{u_{i-1}^2}{4}}{u_{i-1}\left(1 - \frac{u_{i-1}}{2} + \frac{u_{i-1}^2}{8}\right)} \\
&(u_{i-1} \leq 1) > 1 \\
\Longrightarrow\quad & \alpha_{i+1}^n > \alpha_i^n \\
\Longrightarrow\quad & \max \alpha_i^n = \alpha_n^n = \frac{u_{n-1}}{2} \leq \frac{2}{n+2} \qquad (38)
\end{aligned}$$

Accordingly $\sum_{i=1}^n (\alpha_i^n)^2$ can be bounded by

$$\sum_{i=1}^n (\alpha_i^n)^2 \leq n(\alpha_n^n)^2 \leq \frac{4n}{(n+2)^2} \leq \frac{4}{n+2} \qquad (39)$$

Eventually, combining (34) (37), (38) and (39) together yields the proposed high probability bounds (33).

■

By definition $\tilde{G}$ and σ are both upper bounded by G. And if Γ_f is undefined, by combining strong convexity condition $\mathbb{B}_f(x_1||\arg\min f(.)) = f(x_1) - \min f(.) \geq \frac{\lambda}{2}||x_1 - \arg\min f(.)||^2$ and (31) together we can still set

$$\Gamma_f = \left\{ ||. - x_1||^2 \leq \frac{G^2}{\lambda^2} \right\}$$

such that $D = \frac{2G}{\lambda}$, while $\arg\min f(.)$ is always included in Γ_f. Consequently, even in worst cases (32) can be easily superseded by any of (33), in which η decreases exponentially with t instead of inverse proportionally. In most applications both $\tilde{G}$ and σ can be much smaller than G, and σ can be further reduced if each $\tilde{\omega}(x_i)$ is estimated from averaging over several stochastic oracles provided simultaneously by a parallel/distributed system.

V. Discussion

In this article we proposed algorithm 1, a first-order algorithm for stochastic strongly convex optimization that asymptotically outperforms all state-of-the-art algorithms by four times, achieving less than S suboptimality using only $\frac{2G^2}{\lambda S} - 3$ iterations and stochastic oracles in average. Theoretically algorithm 1 can be generalized to strongly convex functions w.r.t. arbitrary norms using technique proposed in [5], and a slightly different analysis can be used to find optimal methods for strongly smooth (a.k.a. gradient lipschitz continuous or g.l.c.) functions, but we will leave them to further investigations. We do not know if this algorithm is optimal and unimprovable, nor do we know if higher-order algorithms can be discovered using similar analysis. There are several loose ends we may possibly fail to scrutinize, clearly, the most likely one is that we assume:

$$\max_f S(y) = \max_f \{f(y) - \min f(.)\} \leq \max_f f(y) - \min_f \min f(.)$$

However in fact, there is no case $\arg\max_f f(y) = \arg\min_f \min f(.) \quad \forall y \in \Gamma_f$, so this bound is still far from unimprovable. Another possible one is that we do not know how to bound $\frac{\lambda}{2}\sum_{i=1}^n \alpha_i^n ||x_i - y_n||^2$ by optimizing x_n and α_n^n, so it is isolated from (23) and never participate in parameter optimization of (27), but actually it can be decomposed into:

$$\begin{aligned}
\sum_{i=1}^{n+1} \alpha_i^{n+1} ||x_i - y_{n+1}||^2 &= \min \sum_{i=1}^{n+1} \alpha_i^{n+1} ||x_i - .||^2 \\
&= \sum_{i=1}^{n+1} \alpha_i^{n+1} ||x_i - y_n||^2 - \frac{1}{2\lambda} || \bigtriangledown_{y_n} \left\{ \sum_{i=1}^{n+1} \alpha_i ||x_i - y_n||^2 \right\} ||^2 \\
&= \sum_{i=1}^{n} \alpha_i^{n+1} ||x_i - y_n||^2 + \alpha_{n+1} \left[||x_{n+1} - y_n||^2 \right] - \frac{(\alpha_{n+1}^{n+1})^2}{2\lambda} ||y_n - x_{n+1}||^2 \\
&= (1 - \alpha_{n+1}^{n+1}) \left[\sum_{i=1}^{n} \alpha_i^n ||x_i - y_n||^2 \right] + \left[\alpha_{n+1} - \frac{(\alpha_{n+1}^{n+1})^2}{2\lambda} \right] ||y_n - x_{n+1}||^2
\end{aligned}$$

such that $\left[\alpha_{n+1} - \frac{(\alpha_{n+1}^{n+1})^2}{2\lambda}\right] ||y_n - x_{n+1}||^2$ can be added into the right side of (27), unfortunately, we still do not know how to bound it, but it may be proved to be useful in some alternative problem settings (e.g. in optimization of strongly smooth functions).

Most important, if $f(.)$ is λ-strongly convex and each $\tilde{f}_i(.)$ can be revealed completely by each oracle (instead of only its first-order information), then the principle of empirical risk minimization (ERM):

$$y_n = \arg\min \sum_{i=1}^{n} \tilde{f}_i(.)$$

easily outperforms all state-of-the-art stochastic methods by yielding the best-ever rate of convergence $\frac{\sigma^2}{2\lambda n}$ [7], and is still more than four times faster than algorithm 1 (through this is already very close for a first-order method). This immediately raises the question: how do we close this gap? and if first-order methods are not able to do so, how much extra information of each $\tilde{f}_i(.)$ is required to reduce it? We believe that solutions to these long term problems are vital in construction of very large scale predictors in computational learning, but we are still far from getting any of them.

Acknowledgements

We want to thank Dr Francesco Orabona and Professor Nathan Srebro for their critical comments on this manuscript. We also want to thank Maren Sievers, Stefan Rothschuh, Raouf Rusmaully and Li He for their proof reading and/or technical comments on an C++ implementation of this method.

References

[1] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 161–168. NIPS Foundation (http://books.nips.cc), 2008.

[2] V. Franc and S. Sonneburg. Optimized cutting plane algorithm for large-scale risk minimization. *Journal of Machine Learning Research*, 10:2157–2232, 2009.

[3] E. Hazan and S Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

[4] T. Joachims. Training linear svms in linear time. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*, pages 217–226. ACM, 2006.

[5] A. Juditski and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex. *Manuscript*, 1:1, August 2010.

[6] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, pages 807–814. ACM, 2007.

[7] K. Sridharan, N. Srebro, and S. Shalev-Shwartz. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, volume 22, pages 1545–1552. NIPS Foundation (http://books.nips.cc), 2008.

[8] C.H. Teo, SVN Vishwanathan, A. Smola, and Q.V. Le. Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, 1:1–55, 2009.